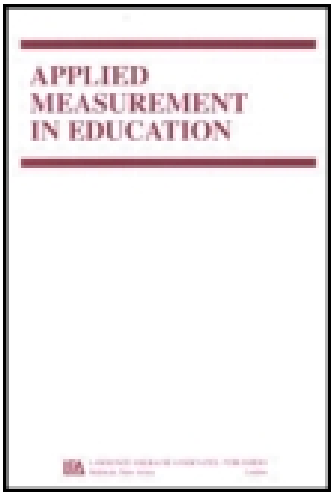

Applied Measurement in Education
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hame20

Examinee Non-Effort on Contextualized and Non-Contextualized Mathematics Items in Large-Scale Assessments

Daniel Van Nijlen[a] & Rianne Janssen[a]
[a] KU Leuven, Centre for Educational Effectiveness and Evaluation, Leuven, Belgium
Published online: 03 Jan 2015.

To cite this article: Daniel Van Nijlen & Rianne Janssen (2015) Examinee Non-Effort on Contextualized and Non-Contextualized Mathematics Items in Large-Scale Assessments, Applied Measurement in Education, 28:1, 68-84, DOI: 10.1080/08957347.2014.973559

To link to this article: http://dx.doi.org/10.1080/08957347.2014.973559

Applied Measurement in Education, 28: 68–84, 2015

ISSN: 0895-7347 print/1532-4818 online
DOI: 10.1080/08957347.2014.973559

Routledge
Taylor & Francis Group

Examinee Non-Effort on Contextualized and Non-Contextualized Mathematics Items in Large-Scale Assessments

Daniel Van Nijlen and Rianne Janssen

KU Leuven, Centre for Educational Effectiveness and Evaluation, Leuven, Belgium

In this study it is investigated to what extent contextualized and non-contextualized mathematics test items have a differential impact on examinee effort. Mixture item response theory (IRT) models are applied to two subsets of items from a national assessment on mathematics in the second grade of the pre-vocational track in secondary education in Flanders. One subset focused on elementary arithmetic and consisted of non-contextualized items. Another subset of contextualized items focused on the application of arithmetic in authentic problem-solving situations. Results indicate that differential performance on the subsets is to a large extent due to test effort. The non-contextualized items appear to be much more susceptible to low examinee effort in low-stakes testing situations. However, subgroups of students can be found with regard to the extent to which they show low effort. One can distinguish a compliant, an underachieving, and a dropout group. Group membership is also linked to relevant background characteristics.

In mathematics education a significant emphasis is put on the use of realistic contexts, both in instruction and testing. Contextualized assignments and test items are often thought to be more motivating and it is argued that they facilitate transfer of what is learned (Cooper & Harries, 2009). Moreover, it is argued that they help to create a more positive attitude toward mathematics. However, contextualized items may also introduce some difficulties (e.g., Boaler, 1993). Contexts need to be translated to a mathematical problem and the solution needs to be re-translated to the original context. One might wonder whether all students will be equally proficient at carrying out this (re-)translation. Moreover, contextualized assignments and test items often put a greater demand on the verbal skills of the student. Language may have an impact on the student performance in content-based areas like science and mathematics (Abedi, 2000). Because of the (generally) larger amount of verbal instruction in contextualized assignments, possibly there is a greater impact of the students' verbal abilities on the performance for contextualized assignments (Abedi & Lord, 2001) and the greater verbal load might also have a negative impact on test motivation for students with lower verbal abilities.

A factor that may introduce construct-irrelevant variance in the context of (low-stakes) large-scale assessments is test motivation (Wise & DeMars, 2005). Because the test bears little or no

Correspondence should be addressed to Daniel Van Nijlen, KU Leuven, Centre for Educational Effectiveness and Evaluation, Dekenstraat 2 PB 3773, Leuven 3000, Belgium. E-mail: daniel.vannijlen@ppw.kuleuven.be

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hame.

consequences for the students not all of them may make all the effort to do a good job. As it is often argued that contextualized items are more motivating, it is worthwhile investigating whether there is any evidence that these items are less prone to low examinee effort. The purpose of the present article is to investigate to what extent contextualized and non-contextualized mathematics test items have a differential impact on examinee effort.

CONTEXTUALIZED ITEMS IN MATHEMATICS EDUCATION

In 1989 the National Council of Teachers of Mathematics (NCTM) introduced their Curriculum and Evaluation standards (NCTM, 1989) in the United States. These NCTM-standards started of the so-called Reform Mathematics (RM) in mathematics instruction. In 2000, these standards were revised and published as the Principles and Standards for School Mathematics (PSSM; NCTM, 2000). These standards were included in the content-area standards for mathematics in many states and in many textbooks (Woodward, 2004). The standards emphasized discovery of knowledge by students and conceptual understanding and de-emphasized things like manual arithmetic. Following the publication of the NCTM standards, in the 1990s a sociocultural framework dominated the understanding of mathematics learning and instruction (Woodward, 2004) and *anchored instruction* (Goldman, Hasselbring, & the Cognition and Technology Group at Vanderbilt, 1997) came to a rise. In anchored instruction the material to be learned is being embedded in authentic problem-solving situations. In current mathematics education the use of contextualized assignments and mathematical problems is widespread (Woodward, 2004).

EXAMINEE EFFORT IN LARGE-SCALE ASSESSMENTS

A prerequisite to get a valid indication of a student's proficiency is that, when tested, the examinee devotes sufficient effort (Wise & DeMars, 2010). However, in large-scale assessments the tests often carry little or no meaning to the test takers (Wise & DeMars, 2005) and from the examinee's perspective the test is "low-stakes" (or even "no-stakes"). This is the case for national assessment programs like the National Assessment of Educational Progress (NAEP) in the United States. The same holds for international testing programs like the Trends in International Mathematics and Science Study (TIMSS) or the Programme for International Students Assessment (PISA). If, because of these low stakes, examinees do not devote full effort, test scores may not be valid indicators of proficiency (Wainer, 1993).

Students may vary in the amount of effort they expend on a low-stakes test (Wise & DeMars, 2005). The existence of different types of test-takers with regard to effort can be grounded in an expectancy-value model of motivation (Wigfield & Eccles, 2000). In this model, effort is determined by the students' expectancy beliefs (influenced by beliefs of their competency and perceptions of the difficulty of the task) and their value beliefs, the beliefs that students hold why they should devote their effort to the task (influenced by, for instance, the intrinsic value and the perceived costs of doing the task). Some students in low-stakes assessments may hold low value beliefs (low intrinsic value and relatively high costs) leading to low effort, despite sufficiently high expectancy beliefs. Other students may see some intrinsic value in doing the test and, in

combination with sufficient expectancy beliefs, may devote sufficient effort. Another group of students may also show low effort because of low value beliefs combined with low expectancy beliefs if the task is perceived as being too difficult to complete successfully.

The influence of test effort can be considered as a form of construct-irrelevant variance (Haladyna & Downing, 2004; Messick, 1989). Systematic, person-specific error is being introduced, resulting in a redefined, biased true score (Lord & Novick, 1968). The study of response patterns and using non-response as an indicator of low examinee effort (DeMars, 2000; Haladyna & Downing, 2004; van Barneveld, 2007) can provide information on potential problems of construct-irrelevant variance in the assessment process.

RESEARCH QUESTIONS

The primary research question of this study is: To what extent do contextualized and non-contextualized mathematics test items have a differential impact on examinee effort? It is investigated whether there is any evidence that contextualized items are less prone to low examinee effort. However, next to this overall effect, it is investigated whether the possible impact is the same for all students or that different types of test-takers can be distinguished. To answer this question one first has to investigate whether students show differential performance on contextualized and non-contextualized mathematics items. If it is found that contextualized items behave in a different way than non-contextualized items, it can be investigated whether there is any indication that the differential test performance could be linked to examinee effort. Additionally, it was explored how the groups of test-takers could be linked to external variables. This helps in gaining further insight into what actually differentiates these groups and it facilitates interpreting the results.

METHOD

Context

In 2002 the first national assessment took place in Flanders.[1] Through national assessments the authorities evaluate at the level of the educational system the mastery of particular sets of attainment targets. These attainment targets are issued for specific educational levels and specify which basic competencies students need to master by the end of the educational level. The national assessments do not have any (direct) consequences for the schools and the students. School level feedback is provided, but students do not receive any individual feedback. The current article presents data from the 2008 national assessment on mathematics in the pre-vocational track of secondary education. As specific attainment targets are set for this track a separate national assessment was organized for these students. The national assessment on mathematics in the general track was administered in 2009.

[1]Flanders is the Dutch-speaking region of Belgium. Authority on education completely belongs to the Flemish government (Ministry of the Flemish Community, 2005).

Participants and Test

The participants in this national assessment were students from the second grade of the pre-vocational track of secondary education (grade 8) in Flanders. In secondary education the Flemish educational system can be described as tracked and non-comprehensive. About 20% of the students in grade 8 are in this pre-vocational track, while the other 80% are in the general, academic track. The second grade of the pre-vocational track prepares the students either for a vocational education that is not aimed at pursuing higher education or for part-time education (Ministry of the Flemish Community, 2005).

A stratified sample of schools was drawn. Three stratification variables were used: school type, educational network, and urbanization level of the region. All students in the school who were in the second grade of the pre-vocational track participated. The total sample consisted of 5,714 students from 195 schools.

The test consisted of 14 subsets of items, each pertaining to a certain topic and was administered in a blocked incomplete design with 15 booklets, each containing five subsets of items. All items were scored dichotomously. The current article focuses on two subsets of items, both on arithmetic. One subset of 28 items focused on elementary arithmetic (addition, subtraction, multiplication, and long division) and consisted of non-contextualized items for which the use of a calculator was not allowed. Another subset of 28 contextualized items focused on the application of the same main operations in authentic problem-solving situations (functional arithmetic). However, because of the context students sometimes had to perform some additional steps like the meaningful rounding of the result. Also, for some items several operations had to be combined. The use of a calculator was allowed as calculator use was considered essential to create authentic problem-solving situations. This choice was in line with the attainment targets and the educational practice in Flanders for this group of students. When in the remainder of the article results on contextualized items are discussed, calculator use is implied.

As the test was administered in a blocked incomplete design, not every student solved items from all subsets. Analyses were performed on data from three out of the 15 booklets that included both subsets of items that were described before. In total data for 1,004 students from 137 schools were included. The booklets were randomly assigned to schools, so students from this group can be considered to be a random subsample of the total sample.

Students were allotted four times 50 minutes to complete a booklet. This time was based on a preceding study where the time needed by the students was registered in each class. The 95th percentile value of these times was used to make an estimate of the time needed. As a consequence, in general, ample time was provided to complete the test. All the items had the same order within a subset, but the subsets did not always have the same position in the booklet. A detailed scheme for the test administration of the three booklets is presented in Figure 1. Sometimes a subset was split with a break in-between. For one booklet the items on elementary arithmetic were administered on two different days. However, all students solved all 28 items of the subsets regardless of the booklet.

Analyses

Differential performance. Differential performance on non-contextualized and contextualized items was evaluated in two ways. First, the stability of the results on both subsets of items

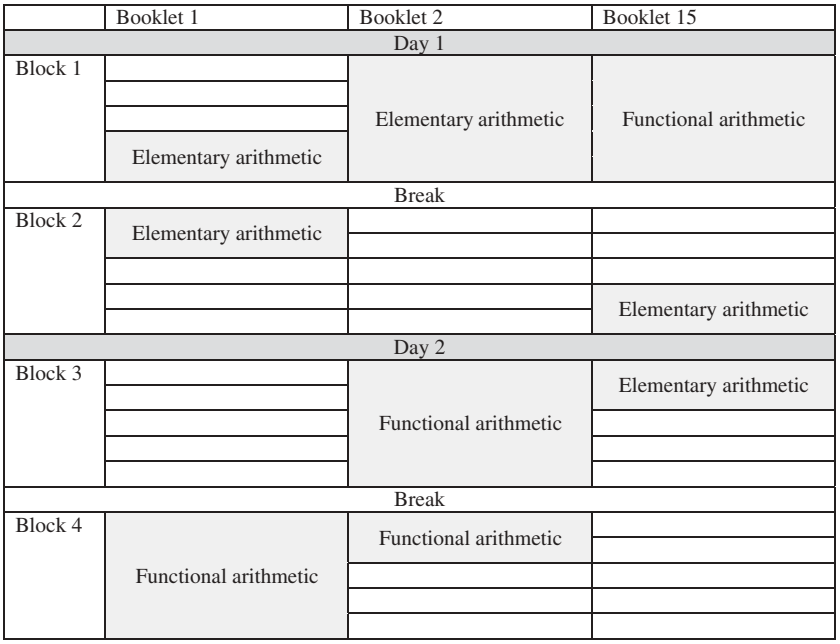

FIGURE 1 Test administration design in four 50-minute test blocks.

was evaluated by correlating the sum scores. Second, response patterns were evaluated by applying mixture IRT models. A differential performance on items for latent subgroups of students would be reflected in different item difficulty parameters for the subgroups. This approach makes reference to the concept of differential item functioning (DIF). While initially DIF was considered to be a nuisance to the measurement properties of a test, it is now considered to provide useful information about the item response process (Zumbo, 2007). However, not always will groups that respond in a different way to items be clear-cut and not always will they be clearly linked to manifest variables. Often the most meaningful grouping is unobserved or latent (Cohen & Bolt, 2005; De Ayala, Kim, Stapleton, & Dayton, 2002; Hoijtink & Notenboom, 2004; Samuelsen, 2005; Webb, Cohen, & Schwanenflugel, 2008).

By applying mixture IRT models to the data it is possible to reveal groups of students that respond in a different way to the items. Mixture models (Fieuws, Spiessens, & Draney, 2004; Lubke & Neale, 2006) divide the respondents in latent classes that each adhere to a different response model. Mixture IRT models combine a latent-class approach with an IRT approach. The model divides respondents into latent classes based on their response pattern, and quantifies performance differences within the latent groups of respondents by means of class-specific IRT scales. Based on the differences in item parameters between the latent classes one can evaluate what constitutes the difference between the groups.

Formal introduction of the mixture IRT model. A specific instance of a mixture IRT model is the mixture Rasch model (Rost, 1990). In this model the respondents are divided into subpopulations for which a certain Rasch measurement scale holds. The mixture Rasch model is an

extension of the Rasch model (Rasch, 1960) that adds a class weight π_g to the equation. This class weight reflects the overall probability of belonging to class g. Also, the ability and the item difficulty (i.e., easiness) parameters are class-specific: θ_{pg} is the ability of person p in class g; β_{ig} is the item easiness parameter for item i in class g.

$$P(y_{ip} = 1) = \sum_{g=1}^{G} \pi_g \frac{\exp(\theta_{pg} + \beta_{ig})}{1 + \exp(\theta_{pg} + \beta_{ig})} \tag{1}$$

Furthermore, we assume $\theta_{pg} \sim N(0, \sigma^2)$ for all g. Since the class weights have to sum to one, π_G is set to $1 - \sum_{g=1}^{G-1} \pi_g$. All IRT (Rasch, 2PL and mixture IRT) models were estimated using LatentGOLD (Vermunt & Magdison, 2000).

Link to background variables. The classification of the students was linked to some background variables. Educational data typically have an hierarchical structure. Students are nested in classes, which in turn are nested in schools. It is important to take this structure into account when modeling the data as it cannot be assumed that the observations within a group are independent, a prerequisite for many traditional statistical techniques. Ignoring this structure would result in an underestimation of the standard errors and, hence, an increased risk of wrongfully finding significant effects (Snijders & Bosker, 2012). Multilevel models take this hierarchical structure into account. They are basically extensions of the standard regression model where some parameters are allowed to vary across higher level units. The multilevel models were estimated with MLwiN (Rasbash, Steele, Browne, & Prosser, 2004).

RESULTS

The average score on the 28 items on elementary arithmetic was 17.97 with a standard deviation of 6.04. The reliability based on Cronbach's alpha was .87. For the 28 items on functional arithmetic the average score was somewhat lower with a value of 15.93. The standard deviation for these items was 4.97. For this scale the reliability was somewhat lower with a value of .78. The correlation between the scores was fairly low with a value of .40. After correction for attenuation it was .48. This means that both subsets cannot be considered to be measuring the same construct.

IRT Models

Both the Akaike Information Criterion (AIC; Akaike, 1974) and the Bayesian Information Criterion (BIC; Schwarz, 1978) were used for model selection (Table 1). These criteria penalize model complexity by taking into account the number of estimated parameters and provide relative information on model fit. The model with the lowest value for the information criterion is preferred.

Both criteria preferred the mixture IRT models over the Rasch and two-parameter logistic (2PL) model (Birnbaum, 1968). Even with six latent classes, AIC still decreased. However, AIC is reported to have a tendency to overestimate the number of latent classes (Nylund, Asparouhov,

TABLE 1
Rasch, 2PL and Mixture Models With One to Six Latent Classes: Loglikelihood (LL), Number of Parameters (Npar), Bayesian Information Criterion (BIC), and Akaike Information Criterion (AIC)

Model	*LL*	*Npar*	*BIC*	*AIC*
Rasch	−32745	57	65884	65604
2PL	−32368	112	65510	64960
Mixed Rasch 2cl	−32049	115	64894	64329
Mixed Rasch 3cl	−31853	173	64902	64052
Mixed Rasch 4cl	−31689	231	64975	63840
Mixed Rasch 5cl	−31550	289	65097	63678
Mixed Rasch 6cl	−31435	347	65269	63565

& Muthén, 2007). Moreover, interpretability of the solution is hampered with a large number of latent classes. The lowest value for BIC was recorded for a solution that distinguished two latent groups of students. The solution with three latent classes was a very close second. BIC clearly started to increase when the number of latent classes exceeded three.

To decide on two or three latent classes, the shift from students between the two solutions was investigated. For each student the probability of belonging to one of the latent classes was calculated and each student was allocated to the latent class with the highest probability. The respective first latent classes of both solutions contain basically the same students with a 92% overlap. The second latent class of the two-class solution is split in two in fairly equal proportions for the three-class solution. The solution with three latent classes can be seen as a more detailed representation of the two-class solution. Given this result and the minor difference in BIC it is decided to continue with the three-class solution. For this solution, the first latent class consists of 50% of the students, the second latent class of 27% and the third of 23% of the students.

Interpretation of Latent Classes Based on Item Difficulty

Each set of item easiness parameters reflects the specificities of the latent class. In the Rasch model there is a one-to-one relationship between the item easiness parameters and the percentage of correct answers. As a consequence the interpretation will be the same whether one uses the parameters or the percentages correct. Therefore, results are presented using the percentages as they are more straightforward to interpret. The results in Figure 2 are presented in the same item-order as they were presented to the examinees. This approach makes it possible to detect order-effects, if present.

The three latent classes showed a different performance pattern on the contextualized and non-contextualized items. The first latent class clearly showed the best performance on the non-contextualized items, but their performance on the contextualized items lagged behind. The second latent class showed a better performance on contextualized items than on non-contextualized items. On top of that, for the non-contextualized items there is a decline in performance on the later items. Their initial performance was close to the performance of the first latent class, but the gap widened as one progressed through the subset of items. For the final items their performance was very close to that of the third latent class. This third latent class had a lower performance on both types of items, but especially for the non-contextualized items. For

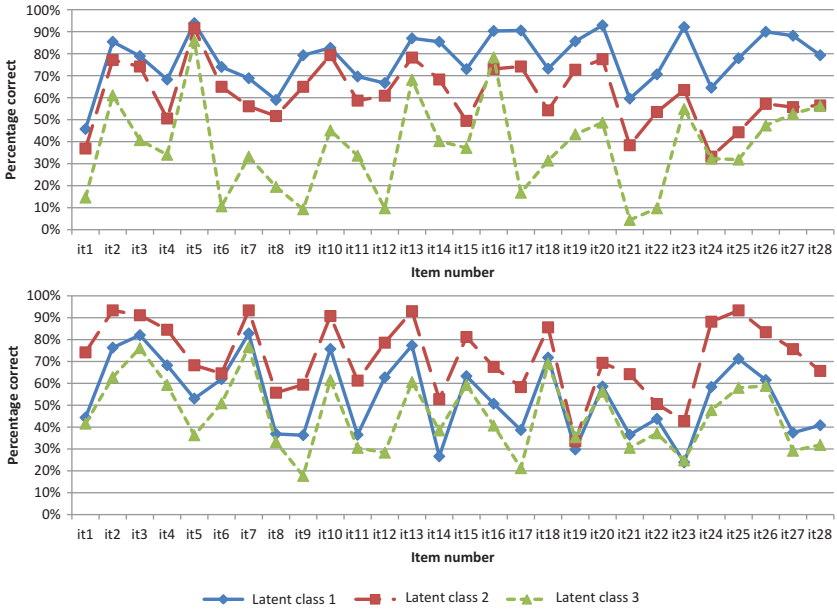

FIGURE 2 Graphical illustration of the percentages correct for the latent classes—non-contextualized items (top panel) and contextualized items (bottom panel).

non-contextualized items concerning long division (items 1, 6, 9, 12, 17, 21, and 22), there was even a further decline in their performance.

Interpretation of Latent Classes Based on Item Omissions

Omitted responses can be considered to be an indicator of low examinee effort (DeMars, 2000; Haladyna & Downing, 2004). For each latent class the percentage omitted responses per item was calculated. Again, results in Figure 3 are presented in the same item-order as they were presented to the examinees.

For the non-contextualized items the omissions showed a very striking pattern. Overall, the level of omitted responses was quite low for the first latent class. There were some minor peaks, mainly for the items pertaining to long division. For the second latent class, the level of item omissions initially was comparable to that of the first latent class, but clearly started to increase midway. For the final items of the subset their level of omissions was comparable to that of the third latent class. The latter group overall had a higher level of omitted responses, but the percentage omissions was very high for the long division items. About half of the time these students did not even make an attempt to solve the item.

Overall, the number of omitted responses was much lower for the contextualized items. Especially for the second latent class this number was very low. Although the number of omissions was somewhat higher for the third latent class, it was very low compared to the number of omissions for the non-contextualized items except for some peaks, all of them for open-ended

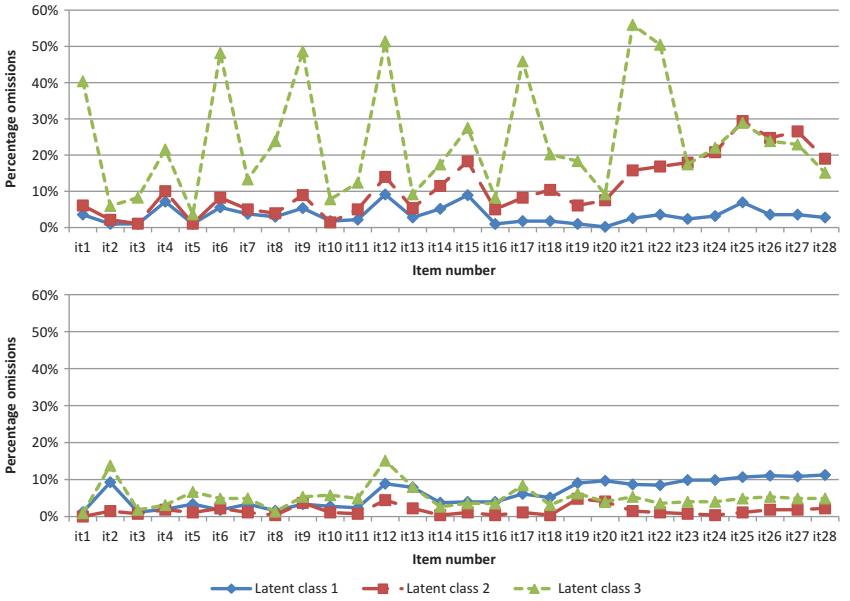

FIGURE 3 Graphical illustration of the omitted item responses for the latent classes—non-contextualized items (top panel) and contextualized items (bottom panel).

questions. For the first latent class there was an order-dependent pattern in the number of omitted responses with an increase for the second half of these items.

Link to Test Booklet

Given that in the three booklets the subsets were administered in different conditions, it was checked to what extent latent class membership was related to the booklet. In Table 2 for each booklet the proportion of students per latent class is presented. There was a significant link between the booklet and latent class membership ($\chi^2 = 48.97$; $p < .001$; $df = 4$). When the non-contextualized items were administered in one block (Booklet 2) students had a considerably higher probability of belonging to the second latent class and a lower chance of belonging to

TABLE 2
Cross-Table of Latent Class Membership and Booklet

	Booklet			
Latent class	*1*	*2*	*15*	*Overall*
1	0.51	0.39	0.63	0.50
2	0.24	0.38	0.19	0.27
3	0.26	0.24	0.18	0.23

the first latent class. When the administration of the non-contextualized items was spread across two days (Booklet 15) students had a higher chance of being classified in the first latent class and a lower probability of belonging to the second latent class.

Latent Class Membership and Background Variables

The dependent variable in the multilevel model was the student's probability of belonging to a latent class. To take into account the covariance between the probabilities of belonging to one of the latent classes a multivariate model was applied. The first latent class was used as a reference group. Based on the estimates for latent class 2 and 3, the parameters and corresponding standard errors for latent class 1 were calculated (Wooldridge, 2009).

Datasets. Analyses were performed on three different datasets. A first aspect included in the first dataset was demographic information on the students (gender, age, and ethnicity), together with information on the presence of learning difficulties (dyslexia and dyscalculia). For ethnicity of the parents a dichotomy was used. The majority of the parents' ethnicity is Western (Western Europe, Canada, and the United States) (82.51%), but for a considerable number of students at least one of the parents was not born in a Western country (17.49%). Also, information on the school's educational network was included. In Flanders two main educational networks can be distinguished. Both of them are publicly funded, but one network is publicly run, while the other is privately run. The students also had to fill out a questionnaire that contained seven statements on their self-assessment regarding mathematics. The students had to evaluate each statement on a 4-point scale indicating the level of agreement. For each student an average score on self-assessment was calculated. For this analysis data were available for 750 students from 118 schools (DATASET 1).

About half of the students also participated in a previous study where they were tested on their performance level for mathematics and reading ability at the beginning of secondary education. A model was estimated with these baseline measures included. In this analysis data for 415 students from 111 schools were included (DATASET 2).

Finally, the latent class membership was linked to the teacher's evaluation of the students' mastery of the attainment targets for mathematics. Teachers were asked to evaluate all students individually as to whether they reached the attainment targets for mathematics. The teacher evaluation was available for 989 students from 137 schools (DATASET 3).

Results multilevel models. First, a three-level model (students, classes, and schools) without explanatory variables was applied to estimate the variance components. The variance at the class level turned out to be not significant, so this level was dropped in the remainder of the analyses. For the two-level model about 10% of the variance in probability of belonging to the first latent class could be attributed to the school attended. For the second latent class this was about 11%. For the third latent class the variance at the school level was considerably lower with only 3%. So, overall the impact of schools on belonging to one of the latent groups was limited, certainly for the third group of students. However, the school level was included in all of the remaining analyses. For each analysis, the intercept reflects the probability of belonging to a latent class for the reference group.

The results for DATASET 1 and 2 are presented in Table 3. Boys primarily have a higher probability of being a member of the second latent class, while having a lower probability of belonging

TABLE 3
Results of the Multivariate Multilevel Analysis—Significant Variables Combined Analysis (DATASET 1) and Baseline Measures (DATASET 2)

	Latent Class 1				*Latent Class 2*				*Latent Class 3*			
Variable	*COEF*	*SE*	*p*	*sign.*	*COEF*	*SE*	*p*	*sign.*	*COEF*	*SE*	*p*	*sign.*
DATASET 1												
Intercept	0.445	0.054			0.145	0.045			0.410	0.040		
Gender (boys)	−0.158	0.036	*0.000*	***	0.133	0.030	*0.000*	***	0.025	0.027	*0.354*	
Grade retention (based on age)												
more than one	0.060	0.060	*0.317*		−0.115	0.051	*0.024*	*	0.055	0.048	*0.252*	
one grade	−0.013	0.032	*0.685*		−0.034	0.027	*0.208*		0.047	0.026	*0.071*	
none*												
Learning difficulties												
Dyslexia	−0.118	0.051	*0.021*	*	0.131	0.043	*0.002*	**	−0.013	0.040	*0.745*	
Dyscalculia	−0.052	0.083	*0.531*		−0.117	0.070	*0.095*		0.168	0.066	*0.011*	*
Ethnicity parents (non-Western)	0.001	0.045	*0.982*		−0.078	0.038	*0.040*	*	0.078	0.035	*0.026*	*
Educational network (publicly run)	−0.084	0.044	*0.056*		−0.010	0.037	*0.787*		0.094	0.029	*0.001*	**
Self-assessment	0.126	0.025	*0.000*	***	0.052	0.021	*0.013*	*	−0.178	0.020	*0.000*	***
DATASET 2												
Intercept	0.593	0.050			0.210	0.036			0.196	0.040		
Gender (boys)	−0.135	0.049	*0.006*	**	0.081	0.036	*0.024*	*	0.054	0.039	*0.166*	
Grade retention (based on age)												
more than one	0.044	0.115	*0.702*		0.052	0.086	*0.545*		−0.097	0.093	*0.297*	
one grade	0.030	0.044	*0.495*		−0.044	0.033	*0.182*		0.014	0.036	*0.697*	
none*												
Learning difficulties												
Dyslexia	−0.151	0.068	*0.026*	*	0.202	0.051	*0.000*	***	−0.051	0.056	*0.362*	
Dyscalculia	0.041	0.115	*0.721*		−0.153	0.086	*0.075*		0.111	0.093	*0.233*	
Ethnicity parents (non-Western)	−0.018	0.069	*0.794*		−0.029	0.052	*0.577*		0.047	0.075	*0.531*	
Educational network (publicly run)	−0.079	0.057	*0.166*		−0.015	0.041	*0.714*		0.094	0.045	*0.037*	*
Baseline measure mathematics	0.005	0.004	*0.211*		0.013	0.003	*0.000*	***	−0.018	0.003	*0.000*	***
Baseline measure Dutch	0.007	0.004	*0.080*		0.004	0.003	*0.182*		−0.011	0.003	*0.000*	***

Note. COEF = parameter value; SE = Standard error; $p = p$-value; *$p < .05$; **$p < .01$; ***$p < .001$.

to the first latent class. Students who repeated more than one grade had a lower probability of being a member of the second latent class. Dyslexic students had a lower probability of belonging to the first group and a higher probability of being a member of the second latent class. For dyscalculia, there clearly was a higher chance of belonging to the third latent class. Students with a non-Western background had a lower probability of belonging to the second latent class and a higher probability of being classified in the third latent class. The self-assessment for mathematics of the students was clearly linked to class membership. Students with a higher self-assessment had a much lower probability of belonging to the third latent class.

For the model that included the baseline measures the power to find significant links was lower because of the considerably smaller size of the sample. Despite the loss of power, overall, patterns for most variables were confirmed. The baseline measures were centered on the total group average so the intercept refers to the probabilities for someone with an average score on both measures. Students with a better performance on the baseline measure for mathematics had a higher probability of being classified in the second group of students. The flipside of this was that they had a lower probability of belonging to the third latent class. There was no link to class membership for the first latent class. Performance on the baseline measure for reading ability was only significantly linked to membership for the third group. Someone with a higher initial score on reading had a lower probability of being classified in this group. This means that someone with a lower baseline score for reading had a higher probability of belonging to this group. As the third latent class performs better on contextualized items than on non-contextualized items this result contradicts the hypothesis that performance on contextualized items would be harder for those who are less proficient in the test language, at least for this specific test.

The results on the link with the teacher's evaluation of the students are presented in Table 4. No other variables were included as this evaluation can also be seen as an outcome variable, and it was thought to be more meaningful to investigate the link with class membership independent of other background variables. Students who, according to their teacher, do not master the attainment targets for mathematics have a much higher chance of being a member of the third latent class. Almost 42% of them belong to this group while for those who do master the attainment targets this is only 15%.

DISCUSSION

The central topic of this study was whether different groups of students could be distinguished in the performance on contextualized and non-contextualized mathematics items and, if this is

TABLE 4
Results of the Multivariate Multilevel Analysis—Teacher Evaluation (DATASET 3)

	Latent Class 1				*Latent Class 2*				*Latent Class 3*			
Variable	*COEF*	*SE*	*p*	*sign.*	*COEF*	*SE*	*p*	*sign.*	*COEF*	*SE*	*p*	*sign.*
Intercept	0.415	0.031			0.165	0.025			0.419	0.023		
Masters AT	0.116	0.032	*0.000*	***	0.149	0.026	*0.000*	***	−0.265	0.026	*0.000*	***

Note. COEF = parameter value; SE = Standard error; $p = p$-value; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

the case, whether differential performance could be interpreted from the perspective of examinee effort. Data were analyzed from a Flemish national assessment of mathematics in the second year of the pre-vocational track in secondary education.

Differential Performance on Non-Contextualized and Contextualized Items

There is a clear indication for a differential performance on the non-contextualized and contextualized items. However, the results also indicate that this is to a large extent due to test effort. The non-contextualized items appear to be much more susceptible to low examinee effort in low-stakes testing situations. For half of the students (latent class 2 and 3) non-response becomes a serious issue for those items. For the contextualized items there is not so much an issue of non-response, although students from the first latent class do show a tendency of increased item non-response when progressing through the test.

Interpretation Latent Classes

Half of the students belong to the first latent class. This group of students performs better on the non-contextualized items. For the non-contextualized items they get the best result of the three latent classes. There is some indication of an order-effect in their performance on the contextualized items. This order-effect clearly shows up for the omitted responses with an increase of omissions for the second half of the items. However, overall the number of omitted responses was fairly low for this group, both on the non-contextualized and the contextualized items. Because of the overall low number of omissions, it could be referred to as a "compliant" group of students. In the expectancy-value model one could say that they combine sufficiently high value beliefs with sufficient expectancy beliefs. Girls had a higher probability of belonging to this group. These students had a higher self-assessment. They also, according to their teachers, more often mastered the attainment targets. Students with dyslexia had a lower probability of belonging to this group. There was no link with their performance in mathematics and reading at the start of secondary education.

The second latent class contains about one quarter of the students. This group shows a better performance on the contextualized items and actually outperforms the first latent class for these items. Their performance on the non-contextualized items deteriorates drastically when progressing through the test. Apparently, this is mainly due to an increasing number of item omissions. This group could be labeled as "underachieving." The fact that their initial performance for the non-contextualized items is comparable to the performance of the first latent class indicates that they actually master these main operations to the same extent as the first group, but underperform on the remainder of the non-contextualized items. For these students one can say that although their expectancy beliefs are sufficiently high their value beliefs are too low to devote sufficient effort for the non-contextualized items. This interpretation is also backed up by information on the background characteristics. These students tend to have a somewhat higher self-assessment and are also considered by the teachers to be mastering the attainment targets. Moreover, these students clearly outperformed the other groups on the baseline measure for mathematics. Some other variables were linked to the membership of this latent class. Boys have a higher probability of belonging to this group, while girls belong to the first group more often. This result is in line with gender differences in examinee effort that were found in some other studies (Eklöf, 2007;

Wise, Kingsbury, Thomason, & Kong, 2004). Also dyslexic students had a higher probability of belonging to this group.

Finally, somewhat less than one quarter of the students belongs to the third latent class. These students show a relatively better performance on the contextualized items, although on both types of items they are outperformed by the other two groups. For the non-contextualized items their performance is clearly lower, but for the contextualized items they almost reach the same level of performance as the first latent class. Overall, they have a high number of omitted responses for the non-contextualized items, but it is extremely high for the items pertaining to long division. The number of omitted responses for the contextualized items is fairly low. These students could be labeled as a "drop-out" group. They perceive themselves (and are perceived by their teachers) as not very good at mathematics and show a severe lack of examinee effort for non-contextualized items, especially for those that seem too difficult or too demanding for them (e.g., division). On the baseline measure for mathematics these students showed a considerably lower performance. Their poor performance on the non-contextualized items appears to be due to low effort, but possibly it also indicates that they experience some difficulties in doing calculations by hand, primarily for the long divisions. One could say that for the non-contextualized items these students combine low value beliefs with low expectancy beliefs.

Test Motivation in Low-Stakes Testing: Lessons for Measurement Practice

To decrease the impact of non-effort different approaches may be needed for the specific groups. The first group of compliant students may not need a specific approach, but in low-stakes situations one still runs the risk that under certain conditions these students will make the shift to the underachieving group: In the test condition where the non-contextualized items were administered in one block more students ended up in the underachieving group. This underachieving group can benefit from measures to prevent non-effort: One can try to increase the value beliefs and decrease the perceived costs. The drop-out group might be very hard to deal with in an assessment situation, as not only their value beliefs are low but their expectancy beliefs for some types of items are very low as well. Several ways to deal with low examinee effort have been proposed (Wise & DeMars, 2005).

On the one hand, proposals have been made with regard to the assessment conditions. These proposals range from raising the stakes of the test, providing incentives or feedback to making the test more intrinsically motivating (Wise & DeMars, 2005; Wise, Pastor, & Kong, 2009). Items that appear mentally too taxing might elicit higher levels of non-effort, but also items that are not challenging or just too boring can elicit this kind of behavior. Also the role of test proctors in examinee effort has been investigated (Lau, Swerdzewski, & Jones, 2009). For this specific test mixing non-contextualized items with other types of items, might have decreased low examinee effort as was already seen when the items were administered in several blocks. On the other hand, proposals have been made to include test-taking motivation into the measurement model (van Barneveld, 2007; Wise & DeMars, 2006). By including test motivation in the model one can get a more accurate measure of both the students' abilities and the measurement properties of the items, provided an explicit measure of test motivation is available.

Mixture IRT models provide a way to check for the presence of specific sub-groups of students in datasets and basically check for the quality of the data. It is shown that in measurement practice

it is also relevant to interpret the presence of subgroups based on item omissions, which implies that registration of these omissions is crucial.

Limitations of and Qualifications to the Present Study

The results show that contextualized items are less prone to problems of test compliance. Even those students that consider themselves to be not very good at mathematics seem to make an effort on the contextualized items, while for the non-contextualized they do not even bother to start working on a lot of the items. However, this difference in examinee effort cannot be solely attributed to the contextualization as calculator use was, in line with the educational practice for students in the pre-vocational track, only allowed for the contextualized items. As a consequence the effects of these two aspects, contextualization and calculator use, cannot be separated, but one could argue that for these students calculator use is crucial to create authentic contexts. In day-to-day practice students of the pre-vocational track are allowed to use calculators for the kind of problems addressed in the contextualized items. This means that it would have been very artificial to administer items on functional arithmetic without a calculator and this would have created other issues of construct-irrelevant variance.

The data pertain to a specific group of students that can be expected to show more problems of scholastic motivation in general. The second year of the pre-vocational track consists of a disproportionate number of grade repeaters and students who have learning difficulties. As a consequence, the issue of non-compliance with non-contextualized items might not be so prominent for other types of students. Paris, Lawton, Turner, and Roth (1991), for instance, found that students tend to take large-scale tests less seriously when growing older. Maybe for younger students the distinction in examinee effort for contextualized and non-contextualized items will not be that prominent.

Another limitation to the study is that no explicit measure of motivation or effort was available. Item non-response is used as an indicator of low examinee effort and motivation is being used as a possible explanation of the classifications of the students. The interpretation based on test motivation is post-hoc and implicitly derived from the results, but supported by different aspects of the analyses.

REFERENCES

Abedi, J. (2000). Standardized achievement tests and English language learners: Psychometric issues. *Educational Assessment*, *8*, 231–257.

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education*, *14*, 219–234.

Akaike, M. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Boaler, J. (1993). The role of contexts in the mathematics classroom: Do they make mathematics more real? *For the Learning of Mathematics*, *13*, 12–17.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, *42*, 133–148.

Cooper, B., & Harries, T. (2009). Realistic contexts, mathematics assessment and social class: lessons for assessment policy from an English research programme. In L. Verschaffel, B. Greer, B., W. Van Dooren, & S. Mukhopadhyay

(Eds.), *Words and worlds: Modelling verbal descriptions of situations* (pp. 93–110). Rotterdam, The Netherlands: Sense Publishers.

De Ayala, R. J., Kim, S. H., Stapleton, L. M., & Dayton, C. M. (2002). Differential item functioning: A mixture distribution conceptualization. *International Journal of Testing, 3–4*, 243–276.

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education*, *13*, 55–77.

Eklöf, H. (2007). Test-taking motivation and mathematics performance in TIMSS 2003. *International Journal of Testing*, *7*, 311–326.

Fieuws, S., Spiessens, B., & Draney, K. (2004). Mixture models. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach*. (pp. 317–340). New York, NY: Springer.

Goldman, S., Hasselbring, T., & the Cognition and Technology Group at Vanderbilt. (1997). Achieving meaningful mathematics literacy for students with learning disabilities. *Journal of Learning Disabilities, 30*, 198–208.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *23*, 17–27.

Hoijtink, H., & Notenboom, A. (2004). Model based clustering of large data sets: Tracing the development of spelling ability. *Psychometrika*, *69*, 481–498.

Lau, A. R., Swerdzewski, P. J., & Jones, A. T. (2009). Proctors matters: Strategies for increasing examinee effort on general education program assessments. *The Journal of General Education*, *58*, 196–217.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lubke, G., & Neale, M. C. (2006). Distinguishing between latent classes and continuous factors: Resolution by maximum likelihood. *Multivariate Behavioral Research*, *41*, 499–532.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement*, (3rd ed., pp. 13–104). New York, NY: American Council on Education and Macmillan.

Ministry of the Flemish Community (2005). *Education in Flanders: A broad view of the Flemish educational landscape*. Brussels, Belgium: Author. Retrieved from http://www.ond.vlaanderen.be/publicaties/eDocs/pdf/107.pdf

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation NCTM standards for school mathematics*. Reston, VA: Author.

National Council of Teachers of Mathematics. (2000). *Principles and NCTM standards for school mathematics*. Reston, VA: Author.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modelling: A Monte Carlo simulation study. *Stuctural Equation Modeling: A Multidisciplinary Journal*, *14*, 535–569.

Paris, S. G., Lawton, T. A., Turner, J. C., & Roth, J. L. (1991). A developmental perspective on standardized achievement testing. *Educational Researcher*, *20*, 2–7.

Rasbash, J., Steele, F., Browne, W., & Prosser, B. (2004). *A user's guide to MLwiN version 2.0*. London, UK: Institute of Education.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*, 271–282.

Samuelsen, K. (2005). *Examining differential item functioning from a latent class perspective* (Doctoral dissertation, University of Maryland, 2005).

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*, second *edition*. London, UK: Sage.

van Barneveld, C. (2007). The effect of examinee motivation on test construction within an IRT framework. *Applied Psychological Measurement*, *31*, 31–46.

Vermunt, J. K., & Magdison, J. (2000). *Latent GOLD*. Belmont, MS: Statistical Innovations.

Wainer, H. (1993). Measurement problems. *Journal of Educational Measurement*, *30*, 1–21.

Webb, M. L., Cohen, A. S., & Schwanenflugel, P. J. (2008). Latent class analysis of differential item functioning on the Peabody Picture Vocabulary Test-III. *Educational and Psychological Measurement*, *68*, 335–351.

Wigfield, A., & Eccles, J. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, *25*, 68–81.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*, 1–17.

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, *43*, 19–38.

Wise, S. L., & DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment*, *15*, 27–41.

Wise, S. L., Kingsbury, G. G., Thomason, J., & Kong, X. J. (2004). *An investigation of motivation filtering in a statewide achievement testing program*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education*, *22*, 185–205.

Woodward, J. (2004). Mathematics education in the United States: Past and present. *Journal of Learning Disabilities*, *37*, 16–31.

Wooldridge, J. M. (2009). *Introductory econometrics: A modern approach* (4th ed.). Cincinnati, OH: South-Western College Publishing.

Zumbo, B. D. (2007). Three generations of DIF analysis: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*, 223–233.